Classification Techniques

Subjects: Transportation
Contributor: Hamed Naseri

Eight classification techniques, including Multi-Layered Perceptron (MLP), Gaussian Naïve Bayes (NB), Logistic Regression (LR), Decision Tree classifier (DT), K-Nearest Neighbor classifier (KNN), Random Forest classifier (RF), Support Vector Machine classifier (SVM), and AdaBoost (AB) were applied to model and predict the CC-SoC. Moreover, these methods were employed to compare the performance of different classifiers and obtain the highest possible accuracy. The classifiers were briefly explained in this section.

Keywords: classification ; feature selection ; climate change

1. Multi-Layered Perceptron

MLP is a deep Artificial Neural Network (ANN) containing more than one hidden layer. ANNs can be employed to model complicated problems in a short time. They are good at nonlinear prediction problems in a reasonable amount of time [1]. An MLP generally includes an input layer, some hidden layers, and an output layer. There are some processing units in each layer, called neurons. All neurons are connected to other neurons by various connection weights (unidirectional connections). The input layer receives the row information, adjusts them, and transfers them to the first hidden layer. The function of the hidden layers is to allocate different weights to each neuron. Then, activation functions are applied to change data representation, and the combination of neuron information and their corresponding weights are transferred to the next hidden layer. Finally, the output layer receives information from the last hidden layer and presents the prediction values or labels [2].

2. Gaussian Naïve Bayes

Gaussian Naïve Bayes (NB) is one of the fastest and most straightforward classification methods. In NB, each sample's posterior probability is maximized during the labels' allocation. NB assumes that the voxel contributions follow a Gaussian distribution, and they are conditionally independent. NB applies a discriminant function for each category. The mentioned function is based on the summation of the squared distances to each classes' centroid weighted by its variance. Then, Bayes' rules are used to calculate the logarithm of the priori probability to train the model. Ultimately, for each testing data sample, the discriminant function is calculated for all classes, and the testing data sample is assigned to a class including the maximum discriminant function value [3].

3. Logistic Regression

Logistic Regression (LR) is a powerful statistical modeling method that has been applied to solve classification problems. LR considers an explanatory variables' set to assess the dichotomous outcome event probability [4]. Dichotomous variables generally denote the occurrence or not of some events. Generally, LR assumes the relationship between the explanatory variables is linear. Thus, LR applies linear decision boundaries while using a non-linear model [5].

4. Decision Tree Classifier

The Decision Tree classifier (DT) was inspired by the shape of trees and their nodes and leaves. DT is easy to understand and interpret. Furthermore, DT easily supports adding new scenarios if introduced, can work as a white-box method, and can be efficient while using an enormous volume of data. Classification rules are mainly modeled based on a set of selections in DT. DT is constituted of decision rules according to optimal feature cut-off thresholds. These thresholds divide each feature into different groups in every leaf node. Then, this process is continued in a hierarchical manner, and at each level, the available samples are divided into different groups based on the splitting criterion [6]. At each step, the current node's branching condition is assessed by splitting criteria. All the mentioned processes are called DT

construction. Subsequently, the pruning process is performed. Pruning is a back forward process that eliminates the additional branches to reduce the computational costs and improve the algorithm's efficiency [7].

5. K-Nearest Neighbor Classifier

K-Nearest Neighbor classifier (KNN) is a black-box classification technique, which has been applied for statistical analysis since the 1970s. KNN is a non-parametric prediction algorithm, and it predicts a sample's label based on the labels of similar samples [8]. KNN plots all samples in a hyper-dimensional space based on their features' values. Afterward, a distance function is utilized, and K nearest samples to the test sample are detected. The test sample's label is the most frequent label in the corresponding K nearest neighbor's label set. Considering a large value for K leads to high running time. Moreover, KNN cannot perform well in the circumstances where more than one frequent label is detected in the K nearest neighbor's label set [9].

6. Random Forest Classifier

Random Forest (RF) is a prediction technique employed for solving regression of classification problems. RF is an ensemble method that combines different DTs to improve prediction accuracy. A particular number of DTs are modeled in the modeling process, and each tree is generated from a random vector. Subsequently, all DT models are run, and the label is determined by considering all DTs' results [10]. Different DT models are run in RF simultaneously, and the majority of class votes determine the predicted label. Research in transport has shown that RF is a powerful method when the problem is large-scale such as an origin-destination survey [11].

7. Support Vector Machine classifier

Support Vector Machine (SVM) is a powerful method used for classification, estimation, and pattern recognition. A set of kernel-based functions are generally applied by SVM to predict class labels in classification problems. Low-dimensional data are converted to high-dimensional vector spaces by nonlinear mapping functions in SVM. As SVM utilizes the theory of structural risk minimization, the over-fitting probability of the problem is reduced [12]. Furthermore, nonlinear complex models can be transformed into simple linear form problems by SVM. Accordingly, SVM can apply linear regression function in a high dimensional space. Consequently, SVM allocates different values of bias and various weights to the model. The SVM model is replaced with a mathematical optimization problem using the principle of structural risk minimization. Afterward, slack variables are added to the new model, and the ultimate prediction model is generated considering fitting error. Ultimately, the optimal solution to the optimization problem is presented as the final classification model [13].

8. AdaBoost

AdaBoost (AB) is an ensemble prediction method that works iteratively. AB combines different weak classifiers in a model to generate an accurate classification method. First, some weak classifiers (sub-classifiers) are generated, and equal weights are assigned to them. Subsequently, the sub-classifiers are trained, and their corresponding error is calculated. Then, the assigned weights are updated based on sub-classifiers' errors, and the updated weights are allocated to sub-classifiers in the next iteration. This iterative process is continued, and ultimately, the class labels are predicted using the results of sub-classifiers and their corresponding weight in the last iteration [14].

References

1. Naseri, H.; Jahanbakhsh, H.; Khezri, K.; Shirzadi Javid, A.A. Toward sustainability in optimizing the fly ash concrete mix ture ingredients by introducing a new prediction algorithm. Environ. Dev. Sustain. 2021.
2. Hasan, K.; Alam, A.; Das, D.; Hossain, E.; Hasan, M. Diabetes Prediction Using Ensembling of Different Machine Learn ing Classifiers. IEEE Access 2020, 8, 76516–76531.
3. Ontivero-Ortega, M.; Lage-Castellanos, A.; Valente, G.; Goebel, R.; Valdes-Sosa, M. Fast Gaussian Naïve Bayes for se archlight classification analysis. NeuroImage 2017, 163, 471–479.
4. Dong, L.; Wesseloo, J.; Potvin, Y.; Li, X. Discrimination of Mine Seismic Events and Blasts Using the Fisher Classifier, Naive Bayesian Classifier and Logistic Regression. Rock Mech. Rock Eng. 2015, 49, 183–211.

5. Hajmeer, M.; Basheer, I. Comparison of logistic regression and neural network-based classifiers for bacterial growth. F ood Microbiol. 2003, 20, 43–55.
6. Suresh, A.; Udendhran, R.; Balamurgan, M. Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers. Soft Comput. 2019, 24, 7947–7953.
7. Rau, C.-S.; Wu, S.-C.; Chien, P.-C.; Kuo, P.-J.; Cheng-Shyuan, R.; Hsieh, H.-Y.; Hsieh, C.-H. Prediction of Mortality in P atients with Isolated Traumatic Subarachnoid Hemorrhage Using a Decision Tree Classifier: A Retrospective Analysis B ased on a Trauma Registry System. Int. J. Environ. Res. Public Health 2017, 14, 1420.
8. Duca, A.L.; Bacciu, C.; Marchetti, A. A K-nearest neighbor classifier for ship route prediction. In Proceedings of the OC EANS 2017—Aberdeen, Aberdeen, UK, 19–22 June 2017; pp. 1–6.
9. Noi, P.T.; Kappas, M. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for L and Cover Classification Using Sentinel-2 Imagery. Sensors 2017, 18, 18.
10. Dogru, N.; Subasi, A. Traffic accident detection using random forest classifier. In Proceedings of the 2018 15th Learnin g and Technology Conference (L&T), Jeddah, Saudi Arabia, 25–26 February 2018.
11. Chapleau, R.; Gaudette, P.; Spurr, T. Application of Machine Learning to Two Large-Sample Household Travel Surveys: A Characterization of Travel Modes. Transp. Res. Rec. J. Transp. Res. Board 2019, 2673, 173–183.
12. Fan, J.; Wang, X.; Zhang, F.; Ma, X.; Wu, L. Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data. J. Clean. Prod. 2020, 248, 119264.
13. Li, L.-L.; Zhao, X.; Tseng, M.-L.; Tan, R.R. Short-term wind power forecasting based on support vector machine with im proved dragonfly algorithm. J. Clean. Prod. 2020, 242, 118447.
14. Hu, G.; Yin, C.; Wan, M.; Zhang, Y.; Fang, Y. Recognition of diseased Pinus trees in UAV images using deep learning a nd AdaBoost classifier. Biosyst. Eng. 2020, 194, 138–151.
